DOCUMENT RESUME

ED 126 016 95 SP 010 209

AUTHOR Follettie, Joseph F.
TITLE Within and Beyond the Formative and the Summative: An Evaluation Perspective for Large-Scale Educational R&D.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO SWRL-PP-23
PUB DATE 16 Feb 73
NOTE 47p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Decision Making; *Educational Development; *Educational Research; Educational Resources; Evaluation Criteria; *Evaluation Methods; *Formative Evaluation; Productivity; School Services; *Summative Evaluation
IDENTIFIERS *Social Indicators

ABSTRACT

This paper schematizes large-scale educational research and development (R&D) as a progression of operations and presents a perspective for evaluating those operations and their outputs. Most perspectives thus far presented for evaluation of educational R&D are oriented to small-scale operations and modest products. Prevailing views of formative and summative evaluation, as developed by Scriven, are analyzed in terms of the state of the art for use of social indicators in isolating first-order and higher-order program effects. Implications of the perspective for educational policy, R&D, and the full-service school are presented. Major dimensions of an evaluation perspective are examined along with organizational and individual roles in improving productivity. Some of the chapters characterize the complex educational product and cause-effect progressions pertinent to complex evaluations. It is concluded that independent evaluation seems required for all evaluations conducted for a sponsor. The best interest of a development organization will be served by independent evaluators working under contract. There will not be any other kind of evaluation of higher-order effects until a system of social indicators is developed, evaluated, and appropriately institutionalized. (SK)

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

Within and Beyond the Formative and the Summative: An Evaluation Perspective for Large-Scale Educational R&D

16 February 1973 Professional Paper 23

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT, 4665 Lampson Avenue, Los Alamitos, Calif. 90720.
Published by SWRI Educational Research and Development, a public agency supported as a regional educational laboratory by funds from the National Institute of Education (NIE), Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position of NIE, and no official endorsement by NIE should be inferred.

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

Professional Paper 23

February 1973

WITHIN AND BEYOND THE FORMATIVE AND THE SUMMATIVE: AN EVALUATION PERSPECTIVE FOR LARGE-SCALE EDUCATIONAL R&D

Joseph F. Follettie

ABSTRACT

Prevailing views of formative and summative evaluation are analyzed in terms of the state-of-the-art for use of social indicators in isolating first-order and higher-order program effects. Implications of the perspective for education policy, R&D, and the full-service school are presented.

CONTENTS

WITHIN AND BEYOND THE FORMATIVE AND THE SUMMATIVE: AN EVALUATION PERSPECTIVE FOR LARGE-SCALE EDUCATIONAL R&D

Joseph F. Follettie

I

INTRODUCTION

IMPLICATIONS OF THE SCALE OF R&D

This paper schematizes large-scale educational R&D as a progression of operations and presents a perspective for evaluating these operations and their outputs. It is contended that effectiveness of educational R&D should increase with its scale. However, the intent is not to recommend a pure-form "big R&D" as an alternative to a pure-form "little R&D." The requirement appears rather to substitute a "mixed" economy for the laissez-faire "little R&D" economy. While this already is happening, its implications for evaluation of educational R&D as yet are underperceived.

Whether by design or oversight, most perspectives thus far presented for evaluation of educational R&D are oriented to small-scale operations and modest products. Perhaps the most extreme manifestation of this tendency yet to appear is the work of Bloom et al. (1971), wherein the classroom teacher becomes a one-person R&D organization who develops and evaluates limited educational routines in the classroom situation, with students simultaneously cast in the roles of learner and guinea pig. Approaches to evaluation that assume a more complex product that is amenable to study within the framework of a multivariate research design also appear in the literature--cf, Scriven (1967), Siegel & Siegel (1967), Stephens (1967), Light & Smith (1970). However, even these more concessionary contributions to large-scale operations addressing complex educational products are largely silent on the important product genesis operations and tend to view the R&D process simply as "develop and evaluate," whether on a one-time or repeated basis.

However dramatic, educationally-referenced efforts that seek to reengineer only one or a few situational characteristics seldom will appreciably ameliorate a prevailing education that suffers in relevance and is underproductive. By definition, educational R&D does not have access to every antecedent underlying educational effects. It cannot secure improved prepartum and postpartum care. Nor can it otherwise in the shorter term appreciably influence the preschooler before he

enters school. Yet an appreciable subset of pertinent antecedents to educational effects are accessible to educational R&D. Every characteristic of educational structure and function potentially is accessible to redesign. Educational R&D probably cannot reach its potential level of effectiveness unless predisposed to address a wide range of accessible determinants of educational characteristics. This R&D should prove complex and so, costly. To insure that the investment is relevant and that it yields a productive return, such R&D will require progressive evaluation over extended time. These considerations suggest appropriateness of what Price (1963) has called "big science" or, strictly speaking, of an analogy to it--"big educational R&D." Yet the evidence is scant that we have thus far broken the mold of "little educational R&D."

"Little educational R&D" features a population of isolated academic entrepreneurs--individuals or small groups--who employ limited perspectives that encompass one or a few of the most accessible situational characteristics. Such activities tend to be independent and uncoordinated. They tend also repetitively to address only that work that the small-scale operator finds easiest and cheapest to do and to sell. The result is a collection of fragments that do not sum to an effective effort. The special contributions that "big educational R&D" can make are those of promulgating more synoptic views and of marshalling the organized engineering efforts that are consonant with these larger views.

The solution to complex social problems does not lie in the direction of pretending that the problems are simple or that piecemeal attacks will suffice. Whether the problem is ecological, socioeconomic, or educational, the reengineering portion of effective responses necessarily requires large-scale R&D. The complex educational products that result from such efforts address an appreciable portion of the pertinent characteristics of an educational system. Such multidimensional designs for educational structure and function that are developed over extended time are illustrated elsewhere (Follettie, 1972). The present paper describes an evaluation perspective for large-scale efforts to create complex educational products.

THE SCALE FOR TYPOLOGICAL COMPLEXITY OF R&D EVALUATIONS

Evaluation theorists have long distinguished between two gross categories of educational R&D evaluative effort. The first, "preliminary evaluation," is conducted when product structure is fluid and so modifiable. Its purpose typically is described as "product improvement." The second, "terminal evaluation," is conducted after the product reaches "final form." Its purpose typically is described in terms of reaching decisions on "product worth." Scriven (1967, p. 43) is generally credited with supplying the terms formative and summative that are now widely used to refer to these categories of evaluation.

Like many dichotomies that are useful in the abstract, the formative-summative distinction has proved ambiguous at more concrete levels.

The terms are used underdiscriminately in the evaluation literature and, too often, with only honorific or pejorative meaning. However, it is possible to ground the formative-summative distinction on a conceptual network that better reflects a characteristic and consistent usage. This paper first sketches this multidimensional network. Then it sets forth categories of evaluation consonant with the conceptual network and technical state-of-the-art for educational R&D. The degree to which the primary categories that emerge "really" reflect the formative and summative labels that are retained throughout the paper may be debated by sophists and Platonists, but is outside the concern of the paper.

As currently used, the formative-summative distinction oversimplifies the progression of decisions that evaluation of educational R&D efforts must serve when the product is complex and costly. The dichotomy was formulated to apply to simple products such as a textbook or other limited educational routine that is to be substituted for one facet of a school whose structure and function are for the most part untouched by the product. Such simple products epitomize the ambitions of "little educational R&D."

R&D programs that are organized to address the full range of issues requiring resolution in education implicate products that are more complex and costly than such simple products. Moreover, the more complex products mature over extended time and so have a greater potential for remaining invisible to the public or to the R&D sponsor during longer periods used to formulate, develop, and operationalize them than do the simple products. For these and other reasons, efforts that yield complex educational products should require more frequent evaluations to serve more decisions than do the plug-in products that ground prevailing evaluation perspectives.

Simple educational products give rise to the view of educational R&D as a linear two-stage process whose first stage, product development, invites classic formative evaluation and whose second stage, product evaluation, invites classic summative evaluation. This view casts all questions relating to genesis of product specifications in limbo and so invites an "evaluator" first to impute genesis decisions to the development organization and second to frame actions that ferret out idiosyncratic biases of the development organization as these show up in the product. In a series of unpublished but widely circulated notes prepared for NIE in 1971-72 to which we will refer extensively, Scriven charts such a course. Only as Scriven's perceptions reference to "little educational R&D" that is grounded on the maxim "Every man for himself" can he be said to be on target.

There is general agreement that all sorts of cut-and-paste operations that serve product improvement objectives may be required during product development. Decisions to modify the developing product to serve such objectives typically are reached on evidence afforded by formative evaluation in its classic sense. These evaluations cannot

occur until the product or some portion of it reaches a form wherein it can be applied to a student so that first-order effects on the student can then be evaluated.

During an earlier product formulation phase of the product development effort (or even antedating it in the sense that the sponsor is able to specify product characteristics), one must specify the domain of product first-order effects (e.g., beginning reading) and the proficiency or behavior dimensions along which first-order effects will be evaluated (e.g., decoding English monosyllables of specified novelty from print to speech, decoding polysyllabic words of specified novelty in light of applicability of morphophonemic rules, supplying appropriate intonation patterns to sentences during decoding). Someone also must specify the educational cost constraints that will apply. If the R&D investment is to be protected, it is necessary also that the development organization specify and the sponsor have evaluated those student transit rates along specified proficiency or behavior dimensions that applicable states-of-the-art warrant (Follettie, 1972). Concerns over monolithic or self-serving "big educational R&D" generally reference to such product formulation activities and particularly to specification of domains and dimensions.[1]

Each product formulation activity is capable of independent evaluation to confirm for the sponsor that the progressing product development operation has social promise and that the sponsor is receiving value on

---

[1]"Big educational R&D" is not incompatible with the proposition that both educational and educational R&D enterprises need to be made more democratic. These enterprises will underserve society to the extent that they are arbitrary and oriented to a self-serving status quo. It appears untenable that we can increase the democracy of these enterprises by giving carte blanche either to enterprise personnel or to parents and students. Somewhere between the extremes of an autocratic establishment that reserves all judgments to itself and an anarchistic one that thrusts all of these judgments on parents and students should lie a social contract that is tenable for educational R&D. The view of a mixed educational R&D economy that seeks seriously to ameliorate profound educational problems must attack a variety of status quo practices. No one group of individuals--professors, R&D personnel, government officials, school personnel, or parents and students--can hope to right the problems of prevailing education while working in isolation. All such groups probably could contribute to a greater extent than they have. We advocate large-scale R&D operations because one cannot hope to bring the different jurisdictions and interests into common effective cause unless they interact within a shared framework that disciplines and focuses the different points of view.

the investment. Each such evaluation either might cause the sponsor to accept progress to that point or to require modification of the formulation as the condition of continued funding. There is little place for such a scheme in classical views on evaluation during product development. Following the classical views, the sponsor cannot hope to shut the barn door until after the horse has escaped. The classic view encourages elitist social planning--whether by funder or developer--that critics of "big educational R&D"--e.g., Atkin & Groteluesche n (1971)--rightfully decry.

When the investment is small, the sponsor will have little incentive to fund the progression of investment insurance evaluations alluded to above. However, when the investment is large, then economic considerations alone should compel that the sponsor sign off or signify displeasure at each of a progression of critical points during product formulation and development. The member evaluations of such a progression each will have summative implications for the sponsor and formative implications for the formulation-development organization(s). However, these summative implications will not be those of classical summative evaluation, which reference to a product in "final form." Rather, these will be implications of a redefined summative category that references to a progression of "summative" entities, where only the last few members of the progression are product entities in the classical summative sense. One tries to excise malignancies early, because the odds are not good that the patient can be saved if these growths are allowed to reach terminal multiplication.

It is also an oversimplification, when the product is costly and complex, to view educational R&D as culminating in a product evaluation stage that permits only classical summative evaluation. One can afford to restrict one's options to accepting or rejecting a cheaply-developed item of any sort. However, we have only to look at the firms that develop, manufacture, and sell the complex systems that power and guide contemporary industry and facilitate modern commerce to see that this range of choices is too narrow when the product is complex and costly. Computer systems typically malperform in minor ways when initially installed in an operating setting. Computer firms would be out of business if they did not have the option of making the initially-malperforming system right following installation. The economics of large investments in educational R&D should compel that product evaluation have summative implications for the sponsor and formative implications for the development organization. However, these formative implications will not be those of classical formative evaluation that references a product not yet ready for evaluation in the operating setting. This is a matter of augmenting Light & Smith's (1970) emphasis on selecting the best products or components with an emphasis on tinkering with good products to make them best buys in the operating setting.

Evaluation of educational R&D for the most part was exclusively summative, in the classical sense, prior to a decade or two ago.

Particularly among educational research faculties, the scales swung toward classical formative evaluation a decade ago, perhaps as a fall-out of the programmed learning movement. Scriven (1967) apparently was the first to see the need for both forms of evaluation. The two warrant equal billing in his earlier views. Hence we must distinguish between the earlier and current Scrivens.

The primary interest of the current Scriven--reflected in his 1971-72 notes--centers on summative evaluation. However, the current Scriven is less interested in the summative evaluations of yore, which addressed first-order product effects, than in summative evaluations dealing with higher-order effects.

The present paper accepts or seeks to extend certain of the current Scriven's views, notably the view that summative evaluations--and conceivably all evaluations--conducted by a development staff risk the biasing of evidence based on conflict of interest. However, the paper argues that Scriven's emphasis on higher-order effects of an educational product is unbalanced and that it is operationally premature in light of the knowledge and technology currently available to support such evaluation. Few would deny that higher-order effects of education are of legitimate concern to society. However, attempts to identify and demonstrate such effects will prove largely ineffective until a system exists that defines baselines against which social cause and effect can be gauged. A later section of the paper considers this matter in light of the views of Bauer (1966) and his associates on the need for social indicators.

II

MAJOR DIMENSIONS OF AN EVALUATION PERSPECTIVE

New perspectives would not be needed if we sought only to enshrine defective prevailing practices. Hence, we need not be too concerned at this point with the well-documented fact that no agency or combination of such has yet emerged to play the central role that effective educational R&D requires. Any candidate agency might use the perspective to be presented as a standard against which it can decide whether it will fish or cut bait. Any other agency might use the perspective as a standard against which to evaluate the rhetorical productions of candidate agencies predisposed to pretend to fish.

Five major dimensions of an evaluation perspective for educational R&D will be discussed. Taken together, these dimensions are meant to be rather exhaustive. However, they show a tendency, as presented, not to be mutually exclusive; some covary to a degree. Where covariation is appreciable, as between the level of product complexity and the scale of R&D, one dimension is counted although both are discussed.

LEVEL OF PRODUCT COMPLEXITY

The product may be simple or complex. Product complexity should appreciably implicate size and complexity of the effort to develop or evaluate the product. Typically, the simpler product will entail "little R&D" and the more complex product "big R&D."

During the era of educational R&D that has heretofore prevailed, it has been customary for one or a few individuals--typically with publisher, governmental, or school district backing--to develop a plug-in or chassis-replacement educational product in consequence of individual perceptions concerning what might sell in the educational market. Such a product may have objectives that are the same as or different from those of a product that, currently in use in the schools, would need be removed to make way for the new product. A product whose objectives are similar to those of a product that it seeks to supplant in the schools is sold on contentions--warranted or not--that the new product for some reason is more attractive than the old. A product whose objectives are novel is sold on contentions that it is more socially relevant than competing products. In either case, the new product typically is viewed as a chassis replacement for an existing product. Its installation typically should minimally disrupt existing structure-function of the schools--or status quo--which, for the most part, the new product will leave intact.

When we view the product so, then the position of Atkin & Grotelueschen (1971) follows that the teacher--like it or not--is the final decision-maker concerning what goes on in the classroom. The position follows because a product that leaves the status quo of edu-

cation appreciably intact cannot hope to do such things as open the door that, closed, transforms the classroom into an inner sanctum. Simple educational products will continue to be developed and marketed. Some of them will undoubtedly contribute to more pertinent and productive education. However, any argument that products must be simple consonant with serving the status quo because--like it or not--the status quo must be served cannot be attuned to the same level of educational disaster that this paper perceives.

The plug-in product orientation is analogous to that for the home equipment enterprise. A host of independent entrepreneural forces mold the home by molding public opinion concerning what constitutes progress--a new gadget (TV), a new wrinkle (color TV), or a new invitation to optimize idleness (an electric can-opener). The process of change is uncoordinated, incremental (if positive), and return-sensitive (whether return is defined on prestige or something more tangible). So it is with simple educational products. These typically do not entail complex, coordinated efforts, are incrementally--rather than comprehensively--oriented to problems of the schools, and are return-sensitive (at best) rather than cost-return-sensitive.

One must acknowledge the views that social programs presently can be designed only as incremental responses to immediate crisis (cf, Braybrooke & Lindblom, 1963) and that educational cost-return concerns are premature. Still, it is presently easy to achieve an increase in comprehensiveness of orientation to educational improvement, if only because level of ambition heretofore has been so low. Quantification of educational cost and return alike pose problems. Nevertheless however cautionary the views of contemporary measurement theorists, school bond elections and the budgeting practices of government alike suggest that the era of educational pigs-in-pokes has ended. Whether the product is simple or complex, it is increasingly likely that product underwriters will want to know what the product will do and at what operating costs. Appeals to prematurity increasingly will fall on deaf ears.

The labels we commonly use to characterize an educational product--treatment, program, product--tend to trivialize the product for complexity. If we take the full-service school as the locus of lower-order product effects, then one may view one or more of the school's services as a complex educational product. A multiyear service then becomes of interest in its entirety as a structure that transits a student from a first-year entry to a last-year exit. Alternatively, the complex product can be viewed as a cross-service entity having functions that are in support of several services of the full-service school. Such a school is not a single model school or experimental school. It is any school that offers a full line of instructional, enrichment, and child care-socialization services.

When we elevate product complexity to that of a service or cross-service component of the full-service school or, ultimately, to the

level of the full-service school itself, the prevailing evaluation view must give way to one that is more sensitive to product formulation steps and to the role of evaluation in investment protection.

It is likely that the other dimensions of an evaluation perspective to be discussed in this section will be differently valued, depending on whether the product to be developed is simple or more complex. Below, these dimensions will be discussed primarily from a standpoint of value-setting implications when the product takes complex form.

LEVEL OF PRODUCT MATURATION

Conventional views on educational product fluidity oversimplify the decision options that a sponsor will find useful when a complex product is to be developed and evaluated. The conventional view is that the sponsor initiates a product development effort that is generally characterized--e.g., as improvement of K-3 reading--and that the sponsor thereafter monitors development operations on an intuitive basis while awaiting product delivery. Such a view neatly partitions evaluation into a formative phase that antedates product delivery and a summative phase that follows product delivery. If this view of sponsorship practice is nearer to fact than to fiction, then practice must be changed. For it makes the sponsor less responsive to technical advice and to educational relevance issues and less responsible to the sponsor's constituency and to development organizations seeking definitive guidance than it should be when costly products that address large educational problems are required.

Consider educational R&D from the standpoint of a sponsor that, interacting over time with an educational product development organization, formulates and develops a desired educational product. The sponsor must first decide which of the organizations that may be available should initiate product formulation. Many consequent decisions of this type can be identified. These decisions all turn on prior sponsor efforts that evaluate capability of organizations based on past performance. We will not further dwell on such decisions here.

Once an organization that will initiate product formulation is identified and oriented to the product domain--presumably on the basis of rather general specifications developed by the sponsor--a product development staff of the organization should proceed to identify product specifications that are consonant with the general guidance, definitive, and acceptable to the sponsor. The effort to formulate definitive product specifications perhaps would reflect formative evaluations of specifications at different points in the effort (or evaluation that, conducted by the development staff, seeks to make the specifications more relevant in macroscopic and microscopic senses--cf, Follettie, 1972). At some point, the effort to formulate product specifications should be ready for independent evaluation that assesses

the effort's structure of product proficiency dimensions for macroscopic and microscopic relevance.

Evaluation of product specifications is relevance evaluation. Conducted long before a definitive product exists, such evaluation cannot hope to extend definitively beyond product first-order effects--that is, the direct effects that most concern a development staff. Evaluation of product specifications could be viewed as formative if product-referenced, in the trivial sense that it seeks to improve the product. However, when the evaluation is viewed as specifications-referenced, it is summative in that it serves a decision to accept or reject the set of product specifications. Conceivably, the evaluation will reference to a standard for social values rather than to a comparative framework that employs the educational status quo as the standard.

It is noteworthy that the same evaluation viewed as summative from the standpoint of a sponsor can be viewed as formative from the standpoint of the development staff that might be required to modify specifications in light of an independent evaluation. The notion that a development staff conducts formative evaluations and an independent evaluation team summative evaluations is an oversimplification. All summative evaluations of complex and expensive educational products have the formative overtone. If the thing evaluated is almost but not quite right, then making it right usually will be economically preferable to starting anew from scratch. Views that contradict this position are persuasive only in the context of trifling investments in educational R&D.

The structure of product proficiency dimensions accepted by the sponsor as characterizing the domain of apt first-order effects, it becomes necessary to place contractual standards on the development staff concerning the extent to which a student will be transited over these dimensions. These standards can be viewed either as criterion proficiency standards, where the investment in operating costs of the school and in student time is specified, or as cost-referenced transit rate standards. The distinction is only terminological. Transit rate terminology is used here.

Product specifications might reveal what proficiency levels should be taken as entry values and indicate upper bounds for student and school contributions to the costs of transiting students across the set of product proficiency dimensions. The development staff should read the applicable states-of-the-art in the context of specified school operating costs and, if available, the experience of prevailing education regarding a comparable existing product. In consequence, it should reach guesstimates that are preliminary transit rate specifications for the product. Since the specified student population should prove heterogeneous both for entry skills and for transit rates, transit rate specifications should reflect both central tendency and dispersion statistics. At some point in such a development effort, the transit specifications should be ready for independent evaluation that judges

preliminary transit rate values against applicable states-of-the-art. This evaluation serves a sponsor decision to accept, reject, or require modification of transit specifications. Hence, the evaluation is summative when specifications-referenced. Needless to say, transit rate specifications for a new product should exceed those that characterize a corresponding prevailing educational product. However, simply exceeding the status quo seems less desirable than exceeding it to the extent that cost-constrained exploitation of applicable states-of-the-art makes possible. Both the development staff effort to produce preliminary transit rate specifications and the independent evaluation of these specifications necessarily will be intuitively based. The purpose of evaluation in this instance is to discourage the development staff from making its work too easy by referencing its effort to an appreciably underproductive status quo.

Preliminary transit rate specifications accepted by the sponsor as consonant with product operating cost provisions and power of applicable states-of-the-art, advanced development should ensue.

During advanced product development, limited tryouts of facets of the developing product will occur. Conducted by the development staff, these tryouts provide the earliest empirical basis for deducing product transit rate characteristics. They form a progression of formative evaluations and modifications that culminate in development of a product that a) is characterized by empirically-based provisional transit rate specifications and b) is ready for full-scale tryout. The sponsor's decision to have a full-scale tryout should stem from evidence, gained during the limited tryouts, that the product is promising for educational productivity. This promise is reflected in provisional transit rate specifications.[2]

---

[2]Findings obtained during limited tryouts condition a decision to have a full-scale tryout, which may be costly. If one views these cut-and-paste-serving tryouts as an informal series that terminates on definitive tryouts for isolated portions of the educational product, then entertainably these terminal members of the series should be viewed as summative evaluations conducted by independent evaluation teams. A possible compromise between terminal limited tryouts conducted exclusively by a development staff that risks conflict of interest and an evaluation team that does not is for the development staff to conduct such evaluations with a technical representative of the sponsor monitoring these evaluations closely. The ultimate extension of this point of view treats data-collection requirements generated by all scientists, engineers, and other interested individuals as subject to conflicts of interest that may, at minimum, distort perception and so bias findings. There is something to be said for this, and the advocates of single-blind and double-blind studies have said it. However, at some point in the effort to eliminate conflict of interest, practical considerations intrude, and one is forced either to accept some capacity for honest appraisal or to create an unmanageable system whereby the police who police the police are themselves policed, ad infinitum.

Independent evaluation during a full-scale tryout establishes tenability of provisional transit rate specifications and serves an agency decision to install the product on a probationary basis--probationary installation. The full-scale tryout is the first of a series of whole-product-referenced summative evaluations. However, its findings might suggest limited product modifications that should occur prior to probationary installation. Moreover, because the full-scale tryout situation will not be isomorphic with the operating-school situation that the product is designed to accommodate, findings might suggest how transit rate specifications should be modified to adjusted provisional transit rate specifications, which will be used during the earliest portion of the probationary installation period to evaluate both product and performance of the schools with regard to the product (see Follettie, 1972). Thus, when viewed from the standpoint of a product development staff, summative evaluation again takes on formative evaluation coloration.

Development staffs and educational evaluation theorists alike have tended to view the total development effort reported up to this point as one to which formative evaluation is applicable but not summative evaluation. One can understand how this view could arise in the climate of an unregulated free R&D market and its dictum that any notion is a good one that sells (even for a few seasons). However, the continuation of this orientation to evaluation when costly complex educational products are to be developed promises to be much too expensive and wasteful to perpetuate. A sponsor should not take an extended costly ride over the route sketched above without assuring itself along the way that social need is being served and that early promise is maturing into something more tangible. Scriven is correct that verbal evidence supplied by a product development organization will not always be disinterested. However, his responses to the problem of conflict of interest seem half measures at best. His outside or goal-free formative evaluation mislabels a progression of evaluations that always can be viewed as summative if properly referenced.

The alternative view presented above saddles the sponsor with responsibilities that have not heretofore been acknowledged. The view of a progression of summative evaluations throughout product development requires any sponsor that acts as the central nervous system for "big educational R&D" operations to lead, whereas candidates to sponsorship roles heretofore have been content to advise and consent.

Additional and important other product development and evaluation efforts lie beyond probationary installation--or within a probationary operation period. Extended products necessitate that the probationary operation period be extended. The customary view of the formative-summative dichotomy entails viewing the period as one wherein only summative evaluation occurs. It is reasonable that first-order product effects should be definitively summatively evaluated during the period. However, when the product is complex and extended, more should occur during probationary operation than classical summative evaluation.

The probationary operation period of an R&D sequence for complex educational products is a hitherto unrecognized necessity. Its analogue is to be seen in all large-scale operations that yield complex artifacts and systems. We explore the period first in the tidier world of commerce.

The manufacturer of complex hardware systems--e.g., computer systems--would not be in business long if not allowed, during a post-installation period, to do whatever is required to bring the system up to contract specifications. The hardware system manufacturer is contractually bound to develop a system that works as well in the operational setting as it does in the factory. The contract is not fulfilled when the product is delivered to a receiving room or installed in an operating room. Only when its first-order effects are demonstrated in the operational setting using inputs that characterize the setting is the contract fulfilled. At that time the manufacturer secures buyer acceptance of the product.

The manufacturer first demonstrates in the factory setting that the system performs according to contract specifications. This test is analogous to the full-scale tryout of an educational product in that first-order effects are evaluated in a setting that is similar to but not identical to the operating setting. The decision to install the system follows from favorable findings obtained in the factory setting. During a probationary period beginning with installation and ending with buyer acceptance, the manufacturer makes whatever adjustments may be required to cause the system to achieve contracted first-order effects in the operating setting. If no problems show up, the probationary period may be quite short. However, if initial evaluation reveals substandard system performance, then its cause must be identified and corrective action taken. It is possible that a series of correction-evaluation routines will be required--a servomechanistic process that culminates when contracted performance is achieved in the operating setting. Perhaps there are instances when costly hardware proves incapable of adjustment to contract specifications in the operating setting. In that case, the manufacturer either accepts the view that the entire effort must be written off or returns the system to the factory for further development. Usually, the system will prove capable of adjustment to contract specifications in the operating setting. That is, in time most such systems that leave the factory will be accepted by buyers. Complex educational products warrant similar treatment during a probationary period and should also in time prove acceptable to buyers in light of performance in the operating setting.

Summative evaluation traditionally has signified hands-off evaluation of the educational product's lower-order effects during the period we call the probationary period. The typical allegation is that an evaluation in the operating setting during a probationary period has no other purpose than to indicate that the product is or is not a good buy. This view may be economically tolerable when the product is on

the order of a textbook. Textbooks are rather cheaply developed and some other textbook always is offstage awaiting its turn when a given textbook fails. However, we should not allow expensive educational products to reach advanced development unless they promise to deliver desired lower-order effects, should not allow their general distribution and use until a full-scale tryout strongly suggests that they will perform up to contract specifications in the operating setting, and should not quit on them in the operating setting when minor adjustments will cause them to perform according to contract specifications. One cannot revive the hopelessly dead in the operating setting. However, in well-managed educational R&D, few products that eventually must be written off ever will reach installation. For all others, the probationary period is conceived here as one that insures that promising investments always will be salvaged. According to this view, the product ceases to be fluid only when it performs consonant with buyer acceptance.

As educational products increase in complexity and cost, it will become increasingly necessary to view an initial decision to install a product in the schools as probationary. Evaluation teams that are independent of the product development effort then might evaluate lower-order effects of the product, with findings fed to the development staff for corrective action when product performance falls below contract specifications. With buyer acceptance, the product reaches a form that can be considered final until advances in applicable states-of-the-art, changes in taste, or evidence of undesirable longer-term effects necessitates that the product be modified or supplanted.

The standards on which absolute evaluation during the probationary period could be predicated themselves will evolve as a progression whose first set consists of adjusted provisional transit rate standards and whose last set consists of definitively stable transit rate standards. The first of these sets stems from a full-scale tryout. The modifier "adjusted" is used because the empirical evidence that the tryout affords will be based on a situation that differs in several forseeable respects. The modifier reflects application of a guesstimation process to the tryout findings. The adjusted standards might compensate for the fact that a multiyear service is simultaneously installed in the tryout setting, whereas its design contemplates longitudinal installation. They might also compensate for a product's tendency, under the pressures of parallel development, to employ certain components--e.g., new equipment, new occupational specialties--in prototypic form during the full-scale tryout.

One contemplates a succession of sets of standards to be devised during the probationary period less to serve product evaluation requirements than to confirm the requirements for school personnel. These standards must be fair if they are to have any role in defining and securing performance accountability in the schools. Much technical work remains to do before agreement can be reached on a standards-setting perspective. We have no recourse to doing this work unless we

are willing to evaluate both product and personnel comparatively--the prevailing inapt strategy. If we want to get the best obtainable performance, whether from a product development staff or school personnel, then we must have standards that are set as high as is fair. Provisional or adjusted provisional standards might suffice for product evaluation under certain conditions. However, the transit rates that these standards reflect typically will be lower than the product warrants when installed in the operating situation. Definition of a progression of sets of standards is particularly indicated when the product is a multiyear service, because product performance should improve year by year in the operating setting until all students entering the higher-year levels of the service are graduates of the lower-year levels. We ask the product to deliver improved performance over the years that are required to transit the student from first-year entry to last-year exit. And we ask school personnel--who are a part of the product to the extent that personnel-training routines are effective--to do their share to insure that product performance increases from the first to the nth year of the probationary period for the n-year service. We cannot make these demands within a comparative evaluation framework. We cannot justify them if we treat the problem of standards-setting arbitrarily or oversimply. Whether the task is to evaluate the product fairly or the personnel that the product implicates in a fair and reasonable way, the probationary period cannot be a hands-off period for the product development staff.

A large-scale operation to develop a complex educational product will feature a progression of operations that classify under product formulation, development, and evaluation headings. Each such operation can be evaluated. Acceptance of its output removes some options concerning operations that follow. Products become increasingly mature and decreasingly fluid as they move through such a progression. The product is most fluid during formulation stages and minimally fluid following buyer acceptance.

Scriven asserts a summative evaluative interest in the product prior to buyer acceptance. However, the summative evaluation domain that most concerns Scriven in the NIE notes is that of product higher-order effects, most of which can be expected to show up only in the longer term--that is, years following buyer acceptance. A later section discusses the conditions under which one can hope definitively to evaluate an educational product for higher-order effects. It is not precluded that Scriven's interest in evaluating higher-order effects is "evaluation" in the policy science sense, where baselines and effects alike are speculative entities.

CATEGORY OF DECISION-MAKER

The conventional view of educational R&D appears to distinguish between a closed-hand classical formative evaluation and an opened-mouth

classical summative evaluation. The classical categories of evaluation lead to a view of summative evaluation as decisive rather than informative. The only decision-maker requiring consideration in the classical framework is the buyer, who does or does not buy. This might prove satisfactory when simple products are to be developed, but fails to protect the investment when a costly complex product is to be developed.

Large-scale operations addressing large educational matters require that both sponsor and development organization reach decisions on the basis of open evaluation operations. Evaluation that serves sponsor decisions seeks to establish relevance and worth of the development effort to date. Evaluation that serves development staff decisions seeks to establish where and perhaps how product effects must or might be enhanced. Essentially, the same evaluation findings serve sponsor and development organization categories of decision-makers. However, the sponsor seeks to evaluate the development organization for potential or achieved productivity, whereas the development organization seeks to evaluate the product (which may include a personnel component) for potential or achieved productivity. Inherent in both objectives is the notion of a standard against which the effort will be evaluated. The formative-summative distinction traditionally has been made a function of the level of maturation for simpler educational products. It could alternatively be viewed as covarying with decision-making category.

LEVEL OF INTEREST IN CAUSE-EFFECT

One may be interested in partial or total lower-order effects of a product per se, in lower-order effects of antecedents other than the product, or in higher-order effects of the product in the context of all other antecedents. While a longer-term interest in lower-order effects or even a shorter-term interest in higher-order effects is not precluded, level of interest in cause-effect typically should covary appreciably with product maturation.

Effects have two primary dimensions: locus and time (or delay). Effect locuses are viewed from a standpoint of distance from the product as antecedent and so are defined independently of maturation level for the product. However, there can be an effect in a remote locus only if first-order effects of the product somehow reach the remote locus. When a product operates on a given student to produce first-order effects on the student, it should require increasing time for such effects to work their way out to increasingly remote locuses. To the extent that delay time does not covary with the time scale for product development, level of interest in cause-effect is independent of product maturation.

An issue raised by Scriven in the NIE notes is whether the formative-summative distinction should mirror a distinction between lower- and higher-order product effects or, to use Scriven's terminology, between main effects and side effects. We will argue that Scriven's interest in higher-order effects of educational products is legitimate but that the state-of-the-art for evaluation of social cause and effect must advance appreciably before it becomes possible to evaluate higher-order effects more than intuitively. Intuition sometimes cannot be avoided. Where it is necessary, the sponsor would do well to protect itself against the possibility that its decisions are based on idiosyncratic tendencies of evaluation teams that the sponsor employs.

An issue not raised by Scriven but inherent in the view that higher-order effects are of legitimate concern to the sponsor is the extent to which antecedents other than the product require consideration. A paramount stumbling block to efforts to date to evaluate social cause and effect is that multiple causes lead to multiple effects. If multiple effects are of interest, then it is highly probable that these effects--and particularly the higher-order ones--stem in part from antecedents other than an educational product whose evaluation is of central interest. Order progressions for antecedents and consequents alike must be considered when one seeks to establish cause and effect in a broad social domain. Such progressions for antecedents and consequents that one may associate with a specified educational product are sketched in Section IV of the paper.

CATEGORY OF THE COMPARISON STANDARD

All evaluations involve comparing something with a standard. The standard may be intuitive or explicit, arbitrary or rationally-defined, demanding or undemanding. Comparative-relative standards tend to be explicit, arbitrary, and undemanding. Criterion-referenced or absolute standards tend to be explicit and rationally-defined; they may be demanding or undemanding and should be demanding consonant with operating cost constraints imposed on exploitation of the applicable states-of-the-art. Where levels of effect are dichotomized into first-order and higher-than-first-order, then the standard presently must be comparative-relative when higher-than-first-order effects require evaluation. When first-order effects require evaluation, the standard may be either comparative-relative or absolute.

If one leaves to a development organization or staff all decisions concerning how stringent its criterion-referenced product proficiency levels will be, then it is understandable that some will conclude that absolute evaluation places a less stringent hurdle in the path of the product development staff than comparative evaluation might. If we block this loophole by defining summative evaluations leading to decisions by the sponsor concerning the merit of a development staff's views on criterion-referenced product proficiency dimensions and levels

(or transit rate specifications), then the logic of comparative evaluation becomes less compelling. Comparative evaluation of first-order product effects encourages a product development staff merely to strive to exceed the status quo. It encourages staff to do no more than build a measurably better product when suitably-constrained exploitation of applicable states-of-the-art should yield a currently best-possible product. Comparative evaluation is an invitation to underachievement.

Prior to limited tryouts that administer prototypic portions of the product to students, all effects of a product are only potential. If product formulation operations are progressively evaluated as sketched above, then it should be possible to use absolute standards to evaluate product first-order effects during limited and full-scale tryouts and during probationary operation of the product without risking a conflict-of-interest tendency of staff to ask too little of itself.

Comparative evaluation of social artifacts for first-order effects is the oldest kind of evaluation. Such evaluation appears most apt when, in consequence of using the simple two-stage model for development and evaluation of educational products, no requirement has been set forth for evaluating product formulation or requiring the product to represent a best-possible effort. It is ironic that some now commend such evaluation as epitomizing sound product evaluation. Most products that are just a little better than those that currently prevail in the schools would have to be judged not worth fooling with.

While comparative evaluations of first-order effects might be warranted on occasion, it is difficult--sometimes to the point of impossibility--to define an acceptable comparison study. Proponents of prevailing education tend to take their first-order gains along intangible or fortuitous proficiency dimensions. Critical comparative evaluations tend to require designers who are as wise, forceful, and persuasive as Solomon. Only when we can agree that given prevailing education is as socially relevant as it should be does a straightforward basis for conducting comparative evaluations exist. Only if the prevailing product then is considered "pretty much attuned to suitably cost-constrained applicable states-of-the-art" does it become a tenable standard against which to evaluate an alternative product.

III

SOME BROADER ISSUES

PERSPECTIVES ON HIGHER-ORDER EFFECTS

In the NIE notes, Scriven is primarily concerned with higher-order effects of educational "treatments." These he describes as goal-free (or needs-bound or consumer-oriented) summative evaluations. Such evaluations address side or higher-order effects. However, Scriven underdescribes the effects his side effects terminology subsumes. Nor do Scriven's comments on side effects evaluation go to the heart of the problem concerning how one will establish baselines against which side effects can be detected and gauged. If side effects evaluation is to be more than a purely intuitive exercise, then social indicators must be provided whose time-referenced series of readings taken prior to side effects evaluation form the baseline against which the side effect will be detected and its magnitude established.

The system of general economic indicators has been under development since the 1930s. While a step beyond nothing, all who know the system agree that it is much less than sufficient for predicting economic effects or characterizing economic cause and effect. Economic and educational antecedents enter into a broader domain of social cause and effect. Bauer (1966) and his associates have been working toward a system of social indicators that can be employed in the broader domain of social cause and effect.[3] Scriven apparently does not believe that his interest in educational side effects requires him to address the broader domain. Conversely, Bauer believes that higher-order consequences of large social programs cannot be established unless evaluation is antedated by an operational system of social indicators that provides firm baselines against which higher-order consequences can be detected and gauged. He also believes that it will not be possible to move far beyond evaluation of first-order consequences of a program if one must wait for the program to come into being before the evaluation effort gives attention to the system of social indicators to which the evaluation of higher-order consequences of the program will be baseline-referenced. Thus, he is drawn to the view that the system of social indicators should generally reference to social need, rather than specially reference, through a specified program, to facets of social need.

Bauer distinguishes between short-term second-order consequences whose social indicators might be specially referenced to the program

---

[3]Land (1972) overviews more recent efforts to design systems of social indicators for use in establishing social change.

to be evaluated and longer-term second-order consequences whose social indicators must be generally referenced to social need if evaluation is to be more than an ex post facto footnote to history. Although Bauer's longer-term second-order consequences are here treated as a progression of higher-order effects, the progression accepts Bauer's position. Even considering the perspectives that Scriven builds into the apperceptive masses of his goal-free summative evaluators, his conception of side effects at best reference only to the educational portion of social need--and then only if the evaluator is an educational Leonardo. Hence, Scriven is constrained either to view long-term educational cause and effect as identifiable from within less than the larger framework that is operating on education or to postpone the search until appropriate baselines are established, where the baseline effort is initiated only after the program to be evaluated is identified. One doubts that Scriven would opt for the first of these Hobson's choices. I see no alternative to the second unless the evaluation team is permitted to intuit its baselines--not a very giant step forward.

Bauer distinguishes between the special short-term consequences of programs of an agency such as NASA and general longer-term consequences of these programs in the larger context of all social programs. Bauer's special short-term consequences include first-order effects and more immediate higher-order effects that an agency such as NASA can anticipate. In Bauer's view, evaluation of longer-term consequences of particular programs cannot aspire to be timely unless a system of social indicators that is analogous to the existing system of economic indicators is operational well ahead of the evaluation team's need to evaluate long-term effects of particular programs. His position is that the longer-term effects of a particular program are to be found in an aggregate social accounting system that is responsive over time both to a particular program and to all other programs that represent social action during the period antedating the search for longer-term effects. Bauer and his associates address the problem of creating a social accounting system whose dimensions are sufficiently inclusive to bear upon longer-term evaluation of diverse social programs and whose measures over time provide a basis--quantitative where possible, qualitative where necessary--for explicating evaluation of longer-term effects. Conversely, Scriven either assumes that such a social accounting system presently exists or that ad hoc selection of its pertinent dimensions and consequent collection of baseline data can occur within the timeframe for summative evaluation of a specified program.

Bauer provides a useful first cut on the classification of effects, one that is implicit in Scriven's distinction between main (or lower-order) and side (or higher-order) effects. In addition, Bauer distinguishes between short-term and long-term higher-order effects (or, in his terminology, second-order consequences), thus preliminarily partitioning these effects. Both Bauer and Scriven seem primarily interested in long-term higher-order effects. However, the two are

differently oriented concerning how these effects might be established. Such effects do not fall outside this paper's domain. However, it is likely that progress in capability for evaluating long-term effects will not occur prior to extensive intellectual and dollar investments of the sort described by Bauer and his associates. If this is correct, then summative evaluation of educational services during the next decade or so is not going to do a very convincing job of evaluating long-term higher-order effects. It is not too early to begin trying to characterize such effects more explicitly. However, these efforts in the shorter term should much more contribute to the state-of-the-art that Bauer sketches than to improving the performance of evaluation teams seeking to establish longer-term higher-order effects.

We may speculate--as is common among practitioners of contemporary policy science--concerning how existing programs are affecting society outside the realm of planned or anticipated lower-order effects of these programs.[4] Disciplined speculation concerning such effects probably should be stimulated by a sponsor that seeks to evaluate its investment in educational R&D. While speculation is not evaluation, even in Noah Webster's sense of estimation, we might view the disciplined speculation of a policy science as yielding policy science evaluations that contrast with the evaluations of evaluation science. When we say that the state-of-the-art as yet does not permit convincing evaluation of higher-order effects of an educational product, the reference is to evaluation science evaluation. Policy science evaluation, grounded as it is on a good deal of intuition concerning both inputs and outputs, is not precluded. My suspicion is that the side effects evaluation that Scriven would have a sponsor fund is policy science evaluation for the most part. There is little point in resisting this approach. Until evaluation science state-of-the-art for higher-order effects is appreciably advanced, the weaker policy science approach to evaluation of such effects may be all that is available to those who are concerned with these effects.

---

[4]Contemporary policy science antedates long-time operation of a system of social indicators and so is a system for reaching policy decisions on the basis of an impoverished information base. At its disposal are diverse aggregate social statistics, some permitting disaggregation when the occasion requires. Available input statistics include number of teachers, teacher salaries, teacher-pupil ratios, per student costs, average daily attendance, and student distributions of various sorts--e.g., across a socioeconomic scale. Output statistics tend to be those that are appropriate to evaluating the educational institution as a babysitting service--e.g., years of education and degrees attained. The policy science that might exist several years following installation of a social reporting system such as Land (1972) advocates conceivably will have appreciable scientific power. The policy science discussed here is the one that is presently available, which cannot be better than its information on inputs and outputs.

ORGANIZATIONAL AND INDIVIDUAL ROLES IN IMPROVING PRODUCTIVITY

Scriven (1967) limited the formative-summative distinction to the educational R&D context. Bloom et al. (1971) attempt to bring the distinction into the context of educational practice. The present paper recognizes that a large amount of educational product development currently is done on an individual pre-industrial basis by the same persons who are responsible for rendering the various educational services--the classroom teachers. That classroom teachers develop and use certain materials day-by-day is an invitation that someone was bound to accept to define formative-summative evaluation categories on operations of the classroom teacher. However, if development of a modern mathematics textbook represents "little educational R&D," then what the classroom teacher can do with available resources must be miniscule indeed. There seems little point in confusing the Scriven and Bloom et al. views on formative-summative evaluation, which most clearly share only Scriven's terminology.

Many currently argue that all product development should be done by the classroom teacher. Such arguments ignore both that the classroom can only support "miniscule R&D" and that the larger "little R&D" to which we have referred is inherently limited by comparison with a responsive and responsible "big R&D." In this light we consider again the position eloquently advocated by Atkin and Grotelueschen (1971). From the indisputable premise that the teacher is the final decision-maker concerning what goes on behind the closed classroom door they derive the conclusion that large, organized efforts that place teachers "at the end of a development/innovation line in which they are expected to implement the bright ideas of someone else" must fail. They also are concerned with counteracting an elitist "social planning that assumes that a particularly wise and prestigious group is possessed of an adequate educational vision to warrant investment of our major available resources in an attempt to replicate that vision throughout the countryside." These are separable points of view.

Centralized versus decentralized social planning is a false dichotomy. The Soviet Union has by now proved the dangers of highly-centralized planning in an elitist government bureaucracy. That government even in the United States is underresponsive to the needs and wishes of too many of its citizens in those domains where government's role is paramount is well known. Yet government and associated large segments of private industry continue to grow as we seek to come to grips with complex interrelated antecedents underlying a present imperfect social fabric. In seeking to improve the student's lot within an industrial engineering framework fostered by government, we must somehow avoid bureaucratic tunnel vision and arbitrariness. That does not argue that the entrepreneural model wherein hundreds of thousands of individuals vie for their own personal, uncoordinated pieces of the action--each usually based on a single-dimensional view of problem antecedents--can serve society as well as larger schemes that are

responsive to all of the antecedents (and consequences) that are germane to improving the lot of the student and of society. Whether the entrepreneur is a single professor in a school of education or a classroom teacher, we cannot hope to make education much better than it is so long as we continue to view the problems as resolvable by many thousand uncoordinated organizations of the one-man show variety. That does not argue that an occasional Louis Braille or Sequoyah will not appear from time to time and set large matters straight that great educational industries uniformly misperceive. Nor does it argue that we ever should create the educational situation that deprives the inspired teacher of the opportunity to do much better than an engineered educational service might allow. When a teacher exceeds performance standards for a service, then one should agree with Atkin & Grotelueschen that we try to determine what it is that such a teacher does that leads to such a consequence. However, the Brailles, Sequoyahs, and inspired classroom teachers are irrepressible. Like cream, they will rise to the top if some set of standards that is akin to milk is available for use in comparative evaluation. A mix of effort at both the level of formulating the full-service school and the level of engineering it is in order. It is highly unlikely that such efforts will produce a monolithic educational vision that we "attempt to replicate...throughout the countryside." Rather, they most probably will produce an inventory of designs that, taking into account the sum of the applicable knowledge that is now available, promise to be much more complex than most individuals operating independently ever could hope to achieve.

However strong the teachers' unions grow, it does not appear compelling that society must accept educational tyranny, whether in the classroom or in administrative offices. The closed classroom door to which Atkin & Grotelueschen refer is a barrier that few who labor for remuneration have been allowed to interpose between themselves and those who meet the payroll. Those who defend the closed classroom door on grounds of academic freedom would do well to give equal stress to the fact that privacy can also be used as a license to steal. Accountability remains a fuzzy notion that masks a variety of motivations. Yet we cannot defend two standards--a relaxed one for those who invoke professional mystique and a strigent one for those who do not. It does not seem reasonable that those who earn a living in any professional field should escape provisions of a fair and reasonable accountability standard.

There always will be professional outputs that are truly professional because, lying at or beyond the frontiers of codified knowledge, they represent new discoveries. Most professionals in every field would be out of business quickly if their livelihood depended on their operating at this level more than infrequently. Rather, most professionals are necessarily technicians most of the time. However complex, the technical component of professional work can be specified and evaluated. Any professional effort that goes beyond the specifiable technical work requirements for doctors, lawyers, teachers, or

educational R&D personnel merits special credit. Such effort yields bonuses that go beyond a technical standard.

If we partition professional effort into technical and professional components, then professional carte blanche loses tenability. The dilemma of teachers is that they seek the professional carte blanche that traditionally has been extended to other professional groups at the very time in history when it is becoming clear that this costly privilege must be scaled down in the other groups.

Atkin & Grotelueschen merely advocate a form of educational R&D that is predicated on teacher entrepreneurs. Bloom et al. (1971) provide such a teacher/innovator with the formative-summative evaluation tools that the role requires. One does not quarrel with the tools that Bloom et al. provide. This would be pointless, since the real quarrel is with the premise that the teacher/innovator must be the exclusive force in educational R&D.

IV

CAUSE-EFFECT STRUCTURE

Needless debate concerning what evaluation work needs to be done when--e.g., during the full-scale tryout, the probationary period, or the period following buyer acceptance--can be avoided by differentiating cause-effect progressions sufficiently to show which cause of which effects one might evaluate. This section more concretely characterizes the complex educational product and cause-effect progressions pertinent to complex product evaluation.

SCHOOL SERVICES

The schools provide mandated and elective educational services, using primary structures that are service-specialized and secondary or support structures that apply to two or more services.

Schools dispense several classes of educational service. Three that seem characteristic of the contemporary school are a) instructional services, b) enrichment services, and c) child care-socialization services. Each of these classes subsumes a set of domain-referenced benefits. Reading illustrates a particular instructional service. Its benefits are reading skills. Observation or discussion of a large shopping center illustrates a particular enrichment service. Its benefits are orienting schemas, whether for exemplary shopping centers or for a generalized view of shopping centers. Classroom attentive behavior and behavior in social situations illustrate socialization services. The educational effects of a socialization service should be to maximize the value of instructional and enrichment services.

An instructional service is designed to render students first-order proficient along each of several proficiency dimensions. An illustrative proficiency dimension for reading is decoding printed English monosyllables of specified novelty to speech. The student comes to the instructional service slightly proficient in decoding skill and should leave the service highly proficient. Whether this happens depends on more than the characteristics of an instructional program. It also depends on characteristics of instructional management, the extent to which student and manager receive apt feedback concerning student progress along the decoding proficiency dimension, and other factors--some arising within the service and some outside it.

An enrichment service is designed to broaden or further differentiate the universe schema that a student brings to a specified instructional service or to informal learning. An enrichment service may be prerequisite to one or more instructional services. Alternatively, the service may have terminal intent, as when instruction of the survey or orientation variety is given. While proficiency dimensions of enrich-

ment services have received less attention than those of instructional services, such dimensions are in principle specifiable. They can be defined on scope, differentiation, and organization of conceptual schemas.

A child care service typically has two functions. The first, custodial baby-sitting, apparently has no educational content and will not be further considered. The second is to advance the child for social behaviors that are of concern both inside and outside the school. If evaluation is confined to manifest behaviors--as opposed to internalized antecedents of manifest behaviors--then the behavioral effects of applications of socialization protocols are discernable and quantifiable to the extent that their evaluation against criterion behaviors requires.

If an evaluation team is required to evaluate a particular educational service, then the other services to which the student is exposed provide an intraschool context to evaluation of the particular service. One searches for first-order effects of the particular service in student performance along that service's proficiency dimensions and for second-order effects in student performance along proficiency dimensions of the contextual services.

Certain of the school's central services have educational import. A system for processing and reporting proficiency test data to interested audiences is illustrative. Such a system addresses a variety of the school's educational services and so is denoted a cross-service component of the school. A cross-service component is useful to the extent that it favorably affects student performance along proficiency-behavior dimensions for those services that the component serves. If we define component first-order effects on a field of first-order effects for services served by the component, then designs for evaluating the cross-service component must be more complex than designs for evaluating an educational service.

Establishment of higher-order social cause and effect is in its infancy. Less well recognized, establishment of social cause and effect at lower orders is a more complex business than is typically acknowledged, for the educational service or cross-service component is only one antecedent or determinant of a product's first-order effects. These complications are discussed below, first for effects and then for antecedents.

DIFFERENTIATION OF EFFECTS

Heretofore, the tendency of public agencies controlling educational R&D and of legislatures controlling education has been to constrain the educational services designer or practitioner with regard to a subject matter domain but not with regard to the proficiency dimensions or criterion levels that an educational service will nego-

tiate or attain. We assume below that the designer of an educational service also will be constrained to address certain proficiency dimensions falling in the skills domain and charged to design a service that transits the student to proficiency levels that are as high as applicable states-of-the-art, exploited to the extent that operating cost specifications allow, will permit.

The effects of the educational service, measured along mandated proficiency dimensions, are first-order effects of the service. Noted earlier, these really are so-called first-order effects--as can be shown using appropriate multivariate experimental designs--because they may be inflated or deflated by the operation of antecedents other than the educational service under evaluation. Until we are able to separate effects of the particular service from other antecedents, effects of every order will potentially reflect the play of other antecedents and so will warrant the label "so-called."

A first-order effect of a reading instructional service might occur along a student proficiency dimension for decoding English monosyllables of specified novelty to speech. A first-order effect of a mathematics instructional service might occur along a student proficiency dimension for summing arrays of two-digit numbers of specified array length.

A specified educational service might affect student performance in some other educational service. Thus, first-order student proficiencies resulting from reading instruction might enhance (or interfere with) first-order proficiencies resulting from mathematics instruction. If mathematics instruction is designed to take reading proficiency as prerequisite, then both reading instruction and mathematics instruction should contribute to what typically are regarded as first-order effects of mathematics instruction. Such effects then contain a mathematics-referenced first-order component and a reading-referenced second-order component, or second-order effect. Effective reading instruction might also favorably affect social behaviors in the school. Conversely, effective socialization services might favorably affect reading proficiencies. When the effect of a specified educational service of the school extends to student performance in a second domain of educational service of the school, the second-domain effect is a second-order effect of the first service. The overall second-order effect of a specified educational service is the sum of its second-order effects in all second domains of the school wherein the student receives service. If we seek to get at second-order effects somewhat definitively during a full-scale tryout, then summative evaluation must result that is more complex than what we typically have in mind when considering full-scale tryouts. It is possible to make such evaluation increasingly complex by substituting the cross-service component of a full-service school--e.g., a system that processes proficiency test data and reports proficiency status of the classroom to interested audiences--for the service product.

Second-order effects show up in the same student whose performance reveals first-order effects. Second-order effects are geographically and temporally locused in the school. They may be short-term, referencing to second domains to which the student is introduced concurrently or soon after introduction to a first domain, or they may be long-term, referencing to second domains that the student will not enter for a long while. It is likely that the summative evaluator will give much greater effort to the evaluation of short-term second-order effects than to longer-term second-order effects. For evaluation of longer-term second-order effects poses some of the same data system problems that evaluation of Bauer's longer-term effects does.

Summative evaluation that is concerned with the totality of a school's educational services over year levels need not distinguish between first- and second-order effects as characterized above. Such evaluation is legitimate if its purpose is to support a decision to accept or reject the system of educational services as a unitary package. The trend seems to be in the other direction--to find out what elements of a large package are working well and what elements poorly, so that one can then proceed selectively when installing educational services or directing R&D efforts where most needed. To collapse the distinction is to surrender useful information.

Educational services also may affect student performance or behavior outside the school. In the short term, these services may affect how the student performs as an individual or in a social setting at home or in the community. In the long term, they may affect how the student performs socially and economically in adult life. Such effects of educational services are denoted here third-order effects. Third-order effects differ from second-order effects by occurring outside the school and so at a geographic locus that is more remote from the educational services to which they reference than are second-order effects. However, the primary reason for distinguishing between student performance and behavior in and outside the school is that the dimensions and data underlying evaluation of second-order effects are (or should be) built into the school's educational services, whereas, as with Bauer's long-term effects, the extensive data system that underlies evaluation of third-order effects remains to be designed and placed in operation. To evaluate third-order effects, we first need to decide what performances and behaviors of the student in home and community are pertinent and then secure enough time series data falling along these performance-behavior dimensions so that we can distinguish between baseline states and any changes that may result from introduction of novel educational services or designs.

A student who is affected by a given educational service or set of services might manifest these effects sufficiently in home and community to become himself the immediate determinant of effects on parents and siblings in the home and other individuals with whom he comes into

contact in the community. In the long term, the school's educational services might operate, through the student, on his companions of adult life and on his own children. Such effects are denoted here fourth-order effects. It would be merely repetitive to note the dimensionalization and data-collection problems that serious attempts to evaluate fourth-order effects entail.

One can imagine higher-than-fourth-order effects, wherein those who are fourth-order affected by educational services affect others, who affect others, who affect others, etc. While such remote effects of educational services may seem far-fetched, many of the side effects and second-order consequences that Scriven and Bauer have in mind are societal states that we can only imagine coming into being in consequence of the operation of a contagion model wherein effects of educational services on students are transferred from person to person until a preponderance of the society is "infected." I do not question the legitimacy of evaluation at this level--here denoted post-fourth-order effects. The problem is that we are not yet set up to attempt seriously to evaluate post-fourth-order effects. Bauer appears more clearly than Scriven to understand the prerequisites to evaluation of this sort. One of these prerequisites is several years of data-collecting within the framework of a general (national) social accounting system. For only when one can establish baselines over time does it become possible to establish whether a higher-order effect (of something) has occurred and, if so, in what direction.

Given the framework of a social accounting system, we probably would inherit all of the problems of those who attempt to use the current system of economic indicators to establish cause and effect. That system, as we know, permits an economist to say that something is happening that is unusual. Usually, lots of these things are happening and most of us can say rather explicitly what they are. The problem is that the system does not yet permit economists to speak with one voice (or two, or three) concerning what the antecedents are of the general higher-order effects that the system of economic indicators reveals. I share with Scriven, with Jencks et al. (1972), and others a concern whether education has long-term effects outside the school and, if so, what form these effects take. However, the tools at hand appear insufficient to permit us to evaluate higher-order or longer-term effects of education more than cursorily.

It is likely that we can extensively evaluate first- and second-order effects of specified educational services anytime we choose to allocate the needed funds to such an effort. Third-, fourth-, and post-fourth-order effects (if any) of the school's educational services should tend to occur in measurable amounts only if educational antecedents more nearly correspond to the totality of educational services than to a particular service. Until we can dimensionalize and measure the various antecedents to these higher-order effects--of which the schooling antecedent is only one, the basis will not exist

for fine-grained analysis of cause and effect at higher levels. Rather, the data system and the baselines that it affords will predispose us, like Jencks et al., to search for large generalizations that are based on gross characterizations of antecedents and effect characterizations that result from averaging procedures.

Few probably would subscribe to the proposition that crudity of tools compels complete inaction in the domain of higher-order effects. Still, a mass of recent evidence that is forged from such tools suggests the possibility that we are creating a false understanding of educational higher-order effects when we use contemporary machinery to establish educational cause and effect. The weight of this evidence--cf, Stephens (1967), Jencks et al. (1972)--suggests that nothing that education does much matters, with offstage overtones that perhaps we should accept the null hypothesis in the educational domain. That is, when we sample in somewhat arbitrary ways--consonant with whatever baseline data happens to be available--from the full domains for antecedents and consequences and thereafter relate both antecedents and consequents to a population as average values, the chances are very good that the generalization will be reached that the laws of cause and effect have been repealed in the educational domain. A single event occurring in psychoanalytic space may engender extended traumatic behavior. People may commit suicide or turn to crime if their income is less than half of the national average for a period of time. Certain patterns of events occurring prior to age 8-13 may predispose the child to delinquency and behavior disorder. The entrepreneur's success may predispose him to outstanding effort. Great leaders may inspire. Yet the great crude studies of educational higher-order effects tell us that 12 years of education will neither harm nor help the individual or the larger society. How novel, quaint, and exotic.

Crude evaluations of higher-order effects of education probably will continue to be conducted until the basis exists for finer-grained evaluations. For those who believe that we require summative evaluation at higher levels, the first priority effort should be less to clamor for additional crude evaluations--which inevitably will occur--than to seek to devise, install, and perfect the social accounting system that, envisioned by Bauer and his associates, underlies more effective evaluation at higher levels.

DIFFERENTIATION OF ANTECEDENTS

Where one seeks to evaluate effects of an educational service, then the service itself is a first-order antecedent. If we follow the structure provided above for effects, then earlier and concurrent services that the student has received and of which the manager is aware stand as second-order antecedents to evaluation of effects of a specified service.

The manager is partially programmed by a routine that leads him to understand his role in the service. The manager is a component of the service to the extent that he performs as the routine specifies. He is independent of the service to the extent that he malperforms or transcends provisions of the routine (e.g., by bringing an exceptionally favorable personality or ingenuity to bear). The student also has some characteristics that the service anticipates and some that it does not and so in part is a component of the service and in part independent of the service. Momentary events and longer-term characteristics of the lives of students and managers outside the school give them service-independent characteristics that are denoted here third-order antecedents. A recent crime, a war, a long-term socioeconomic condition, or the company that a student keeps may enter the school as a student-referenced third-order antecedent. Similar events and conditions plus effects of schools of education, in-service training, and exposure to general and professional media enter the school as manager-referenced third-order antecedents. Often it is more effective to directly evaluate social effects in terms of student and manager behaviors and performances that, as third-order antecedents, are brought to the educational service than in more abstract terms--e.g., socioeconomic--referencing to the community.

Social ills find their way into the schools on the shoulders of third parties when the schools are under grave attack by the community or some portion of it. When evaluation occurs in a confrontation climate, then it may be necessary to consider third-party effects that are here denoted fourth-order antecedents.

The fifth-order antecedents that come most quickly to mind are those that illegitimize evaluation of an educational service by asking it to perform under conditions that are contradictory to its design. Thus, funding slashes and countermanding administrative directives that make it impossible to render a service as designed are fifth-order antecedents that transform a service into a caricature of itself. The evaluation team is left then to determine whether the caricature is effective.

All of the antecedents thus far cited are educational antecedents. They enter the schoolhouse through channels or outside channels. Above we implied that education must have some higher-order effects. That is not to argue that a higher-order effect is not also a consequence of antecedents that fall outside the educational domain. Entertainably, every effect that might be of interest to educational evaluation is a joint function of educational and noneducational antecedents. It should be the case that lower-order effects defined narrowly on first- and second-order educational antecedents would be more a function of such antecedents that higher-order effects defined more generally on social need, which should give greater play to higher-order educational antecedents and to noneducational antecedents. Were we to characterize a wide range of noneducational antecedents having shaky

baselines and then use gross study methodology to determine the contribution of each of a totality of antecedents to a higher-order effect defined on social need, it would not be surprising to discover that lower-order educational antecedents are not the only ones that would show up ineffective. Given equally shaky baselines across the board, the grand conclusion should fall out that nothing really matters. To the extent that Jencks et al. can point a finger at anything, they point to an antecedent whose baseline data is not shaky.

When the schools are not under grave attack and when school administrators have the resources and display the determination to operate an educational service as designed, summative evaluation probably needs to consider only first-, second-, and third-order antecedents of lower-order effects. All of these antecedents can be evaluated at the locus of the service. Where the school is under attack, it may be necessary also to consider fourth-order antecedents.

Where the school is not under attack, then so-called first-order effects are a function of first-, second-, and third-order antecedents. Those theorists who are attracted to higher-order effects might pause to consider how tenuous such an enterprise becomes when we are forced to admit that not even first-order effects can adequately be accounted for simply by referencing them to a first-order educational antecedent. It appears warranted that the domain of antecedents will expand--and more than linearly--as one mounts the order scale for effects.

If mountains are there to be climbed, then we will ascend the mountain of summative evaluation. However, the state-of-the-art for evaluation of social cause and effect is such that a practitioner of evaluation cannot hope to get much higher than a first base camp at present. The higher levels should for the most part during the next decade be the province of those whose interests and talents are consonant with advancing state-of-the-art for evaluation of social cause and effect.

EXEMPLARS OF LOWER-ORDER EFFECTS EVALUATIONS

Most references to evaluation thus far made have assumed a first-order antecedent that takes the form of a specified educational service. When the first-order antecedent is such a service, then first-order effects are evaluated in terms of student performance along pertinent dimensions for the service. These dimensions might be enumerated in consequence of the joint efforts of a development staff, a sponsoring public agency, and consultants who are available to both.

Alternatively, the first-order antecedent might be a cross-service component of the full-service school. An example is a system that insures the flow of first-order data for the different services to all interested audiences--e.g., service managers, administrators, and

parents. Such a system might be used to process, organize, and report progress, by individual and class, for each of the school's instructional services. In this event, the first-order effects of all instructional services might be established when the system is in use and when an alternative first-order antecedent replaces the system (e.g., whatever is customary, including the default system whereby no first-order data flows from the classroom concerning any instructional service). The system's first-order effects then are established by comparing it with an alternative (including nil) system for effects on first-order effects of the different services. Positive first-order effects of the system are, of course, further evaluated within a cost-return framework.

When the first-order antecedent is a specified educational service and the second-order antecedent is one or more second-domain services, then the second-order effects of the first-order antecedent show up in the first-order effects of the second-domain services. Where the intent of the first-order antecedent is to supplant a prevailing version of the service whose objectives are to transit the student along identical proficiency dimensions, then second-order effects of the new service relative to the prevailing service can be established comparatively. If the new service has more desirable second-order effects on second-domain services than does the prevailing service, then the first-order effects of second-domain services will be more desirable when the new service is the first-order antecedent than when the prevailing service is. It will tend to be the case that a new service that has an edge in second-order effects but performs in an inferior way for first-order effects will prove unacceptable. Cost-return considerations apply to all examples.

When the first-order antecedent is a cross-service component of the full-service school, evaluation of second-order effects requires a greater investment than when a cross-service component is evaluated for first-order effects or a service is evaluated for second-order effects. The niceties ignored--e.g., random block design--a two-factor factorial design is required when the task is to evaluate second-order effects of a cross-service component. To evaluate such effects, the evaluation design must reflect alternative versions of the cross-service component and, as a minimum, alternative versions of a specified service. The difference between second-domain first-order effects of the two versions of the service for one version of the cross-service component are second-order effects for the new version of the service. The difference between these second-order effects of the two versions of the cross-service component are second-order effects for the new version of the component.

Higher-order educational antecedents ignored, evaluation design increases in complexity as one moves from the service to the cross-service component as a first-order antecedent and from first- to second-order effects. Even when antecedents are viewed narrowly in terms of

first- and second-order domains, the complexity of apt evaluation designs mounts with elevation of interest to higher-order effects. Evaluations of the sort sketched above may be appropriate to the full-scale tryout and to probationary installation-operation antedating buyer acceptance. It is doubtful that evaluations of higher-order effects could occur prior to buyer acceptance of the service, set of services, cross-service component, or set of cross-service components.

There is virtually no way to systematically vary third-order educational antecedents without deliberately reforming the society to conform to provisions of one's experimental design. Antecedents above second-order typically can be varied only in the fortuitous sense of selecting schools in different socioeconomic neighborhoods or having other characteristics defined on demographic central tendencies. When this is done, the characterization of a statistical treatment group tends itself to be no more than a hypothesis concerning what characteristics of the third-order antecedent are pertinent. Moreover, the characterization, if a central tendency, might come near to describing an empty set--with people around it but no one actually there. Problems do beset us when we harken to Scriven's laudatory call to become more ambitious in the educational evaluation domain. Surely in depth we must. In scope the constraints noted above presently preclude much change.

V

ENDS AND MEANS

A DECISION PERSPECTIVE FOR SERVICE PRODUCTIVITY

A multiyear service structure whose function is to transit students along specified outcome dimensions will do so productively to the extent that structure is consonant with productive function. A first general objective of an educatonal R&D program is to create the wherewithal for a service whose theoretical productivity is as high as we could hope to make it in light of applicable states-of-the-art and controlling educational cost constraints. A service's theoretical productivity is what it would do per unit cost if all of its components--and particularly its personnel--perform up to capability. When the service is installed on a probationary basis, its achieved productivity--evident in its performance--should fall below its theoretical productivity for a number of reasons--some referencing to personnel and others to other components of the service. A second general objective is to bring achieved productivity into line with theoretical productivity.

Given the present state-of-the-art, it is necessary to move toward definitive characterization of a service's theoretical productivity concurrently with efforts to optimize the operating service's achieved productivity (see Follettie, 1972). We currently lack the technology to examine personnel, students, and other characteristics of a design-form service and in consequence specify mean and dispersion values for students transiting the service. The operating service must be used to determine the transit rate distribution that the service compels. Perhaps one reason why school personnel characteristics are taken as falling outside the bounds of the educational engineering effort in some accounts is that this view nicely resolves the problem of distinguishing between theoretical and achieved productivity. There is no such problem if we are stuck with whatever performance school personnel care to provide. A second related way to avoid the problem is by requiring summative evaluation to be comparative evaluation. At its best, comparative summative evaluation is predicated on the same level of performance by personnel of the old and new versions of a service. The present paper treats educational service personnel as integral components of service structure and so as having performance capabilities that can be estimated under appropriate empirical conditions. This view reintroduces the problem of distinguishing between theoretical and achieved productivity and makes summative evaluation of lower-order effects more dynamic and difficult than many yet concede.

A scenario sketch may clarify the character of the complexities involved when a multiyear service is to be evaluated for lower-order

effects. Let us imagine that a sponsor charges a development organization with developing the wherewithal for a service (or a version of the service that is alternative to a prevailing version) that will use six instructional years for 30 minutes per instructional day and will transit entering first graders over specified service dimensions as state-of-the-art-optimally as specified operating costs for the service allow. Allow the development staff three years to develop the product to the point where it is ready for a full-scale tryout, the development staff and an independent evaluation team one year for a full-scale tryout conducted under the condition of simultaneous installation across year levels for the service, and the development staff and evaluation team six years for probationary operation wherein the service transits an entering first grader from his point of entry to an exit point that is rate-optimal for the student.

During the full-scale tryout, the evaluation team should be collecting data near-continuously and passing along to the development staff any findings that might be pertinent to modifying the service as a prelude to probationary installation. If the product in tryout form is unpromising, that might be the end of it. If it is promising, then tryout evaluation might yield data--quite possibly incidental to first-order effects evaluation--that suggest how the product might be made more promising still. In this event, there would be little point in the development staff keeping its hands off the product whether in the sense of the installed service under evaluation or the form it will take during probationary operation. Small modifications should be in order.

During probationary operation, we might again think in terms of a defined data flow that the evaluation team monitors and passes along to interested audiences, including the development staff, whose responsibility would be to fine-tune the product to optimize it for productivity.

The setting of productivity standards, the decisions to install on a probationary basis and to accept the product, and evaluation that most aptly serves these decisions could all be made easier if we were to allow the full-scale tryout to use six years--and would become more straightforward still if it were possible where necessary to repeat the six-year tryout. However, unless applicable states-of-the-art are advancing much more slowly than we imagine, a service would be hopelessly outdated before reaching probationary installation with such a generous full-scale tryout period. Such problems do not arise when the product to be evaluated is a simpler one having much shorter theoretical transits. If all involved are willing to be flexible, then the one-year full-scale tryout buys problems that we can afford.

A full-scale tryout inevitably occurs under conditions that are not isomorphic with the designed-for situation. Thus, it may be found

useful during a tryout to minimize personnel malperformance through overtraining and overevaluation or by utilizing R&D personnel in selected positions. It usually will prove necessary to install a multiyear service simultaneously across year levels for tryout purposes. This installation strategy rules out that entrants to higher-year levels will be graduates of the service's lower-year levels, which typically contradicts design specifications for the multiyear service. Hence, first-order effects obtained in the tryout situation will depart somewhat from what would be expected if the product were longitudinally installed in the operating setting and used available personnel trained according to provisions of the service specifications. The purpose of a full-scale tryout, then, is to test the promise of the product in operational use. Evaluation in the tryout setting serves a decision to install or not install the product for a probationary period. The tryout provides a basis for distinguishing between theoretical and achieved productivity under conditions that depart from design specifications, where some of these departures are favorable to service performance and others are not. Definitive standard-setting can only occur during probationary operation.

A decision reached to install the product for a probationary period should usher in summative evaluation that definitely establishes first- and second-order effects. This evaluation effort is in support of a decision to accept or not accept the product as designed until a) the service becomes obsolete or b) apprehended higher-order effects suggest a need to modify or supplant the service. Throughout the probationary period, some staff should be progressively modifying standards defining theoretical productivity toward definitive standards that, representing a fair contract, will characterize achievable productivity.

A decision reached to accept the product might usher in summative evaluation that establishes higher-order effects of the service or of the full-service school. Bauer's views on social cause and effect suggest that summative evaluation at this level would need be general and so defined on broader social need than the school alone ever could address.

The apparent ultimate consequence of Scriven's views on summative evaluation of higher-order effects is that such an evaluation would reference narrowly to a specified simple educational product. Since the educational antecedents to higher-order effects all are pertinent to these effects, the full-service school probably provides a better basis for the antecedent referencing of higher-order effects evaluation than does some one facet of its educational effort. With so many arguing that formal education has no measurable higher-order effects, who could believe that some small part of it could. Policy science may not be dismayed by Scriven's proposition. An empirical science would have to be.

Scriven's "consumer orientation" to summative evaluation should compel him eventually to return to the lower-order effects evaluation domain that the earlier Scriven charted. When costly and complex products are to be evaluated at this level, a host of difficult problems remain to resolve before the work can be considered technically routine. Something like the progression of summative-formative interactions that are sketched above appears necessary because complex educational products perform over extended time. All of this performance is pertinent --both to sponsor decisions and to subordinate decisions by development staffs. Grand summations having no interim import can only occur when the product is of modest proportions. Who much cares when that is the case?

DEVELOPMENT-EVALUATION CONTRACTING PROCEDURE

In the NIE notes, Scriven is centrally concerned with conflicts of interest that may arise if one allows a development staff to summatively evaluate the product developed. That concern is apt. However, Scriven quickly goes from there to the idea of outside evaluation teams to whom the sponsor gives full discretion to decide what the pertinent dimensions of evaluation are at every order level, what order levels are pertinent, etc. Essentially, Scriven advocates giving carte blanche to the evaluation team, presumably on the basis that private such groups with their individual proprietary interest are the nation's best source of consumer protection Ivanhoes (a characteristic that is additional to Leonardoesque proficiencies). Carte blanche threatens every society that gives it to any small group--whether elected, appointed, or self-appointed. Scriven is correct to seek to remove conflicts-of-interest temptations from development staffs. These must be removed from all sources of participation in the R&D enterprise. However, Scriven merely transfers a license to steal from one group to another.

While one can accept the view that others can evaluate a brainchild in a more disinterested way than can its creator, it does not follow that an outsider is more competent to perceive an apt design for evaluation than is an insider. Merton (1972) nicely responds to the view that location outside an operation somehow guarantee's objective purity. According to Merton, "The role of the Outsider no more guarantees emancipation from the myths of a collectivity than the role of the Insider guarantees unfailing insight into its social life and belief-systems." I would make the statement bidirectional. Different points of view are useful because they are predicated on different biasing premises, rather than because of some of them transcend personally-referenced hangups. As with so many macrodimensions of life, the extreme values of solipsism and naive realism tend not to be utilitarian alternatives to a middle ground that makes us all observers who can only to some extent overcome the shackles of personal experience to gain intersubjective views that many can share. I concur with Merton's view that

the search for truth will best be served if insider and outsider interact with each other during the quest. However, when this happens, then insider-outsider terminology becomes less descriptive because we then bring all those with pertinent views in some sense into the same tent. The issue concerning who conducts a specified evaluation (or executes its design) is separable from the issue concerning who designs the evaluation (or the issue concerning how it is designed). When we separate these issues, conflicts of interest are diminished or removed and the stage is set for bringing to bear the different pertinent points of view whose joint consideration insures that a strong evaluation design will be produced.

All interested and qualified parties may participate in the design of summative evaluations--whether classical or in the extended sense sketched earlier. Scriven's notion that an independent evaluation team might, by becoming aware of proficiency dimension specifications advocated by a development staff, in some sense be contaminated, is no more than the notion that some individuals are quite impressionable or easy to dominate. A broadly-based design effort surely will embrace some such individuals. However, the more-likely consequence of a broadly-based effort is a "Tower of Babel" of idiosyncratic points of view that refuse to consider other pertinent points of view. (The problem of development staff idiosyncracy disappears when we agree to extend the summative evaluation concept down to formulation of product specifications.)

Responsibility for executing designed summative evaluations may be discharged according to normal contract-letting and contract-monitoring procedures of the sponsor. The contracting organization should have a track record that indicates competence in areas specified by the design-reflecting contract. The organization should manifest no conflict of interest. Were there a market for contracted evaluations, it is likely that private industry quickly would evidence the required evaluation capability--assuming that comments as are scattered throughout the present paper first sketch the pertinent state-of-the-art.

VI

CONCLUDING REMARKS

Evaluation efforts are separable from development efforts--and increasingly economically separable as product complexity increases. Earliest evaluations address relevance issues--whether intuitively or empirically--and the question concerning how much proposed product specifications exploit applicable states-of-the-art consonant with imposed bounds for operating costs. Later evaluations address product productivity as defined on lower-order and particularly first-order effects and longer-term relevance-productivity as defined on higher-order effects.

Whether we should bend formative-summative evaluation terminology to the description of a formulation-development-postdevelopment progression of evaluations--as I have done--is not an important issue. All such evaluations might be viewed as summative from the standpoint of a responsible sponsor and as formative from the standpoint of a development organization that is charged with modifying the evaluated work. The sponsor and the development organization can be viewed as joint consumers of the same set of evaluative findings. Particularly when the product is complex, findings that are of primary interest to one of these consumers often will prove no less than of secondary interest to the other.

Contracted independent evaluation seems required for all evaluations conducted for a sponsor. Entertainably, the best interest of a development organization also will be served by independent evaluators working under contract. The evaluation organization can no more be given carte blanche concerning what work it will do than can the development organization. Once we extend evaluation down to product formulation activities, the notion that the dimensions for proficiency-behavior evaluation are idiosyncratic inventions of the development organization or staff loses all credibility. At that point, so does Scriven's notion that the evaluation team should have carte blanche.

The two-stage educational R&D perspective that gives rise to the classical view of a formative evaluation period followed by a summative evaluation period is that of "little educational R&D." When the product is viewed as incorporating much or all of the educational structure of the school, brought to bear on an appreciable portion of the school's functions, we reach a large scale level of concern that small diverse R&D efforts cannot effectively address. "Big educational R&D" then becomes appropriate. Such effort need not be monolithic and can be spared this fate if the sponsor is required to act responsively and responsibly and so to involve all communities that have a contribution to make. It is not inevitable that a system of large organized educational R&D strive, like the French Third Republic, to orient each

penciled hand to the same point in a national educational program at a given instant in time. The Soviet Union incurs penalties in important areas because its education is not appropriately balanced along the conformity-independence dimension (cf, Bronfenbrenner, 1970). We can hope to have large-scale educational R&D that is not the pawn of bureaucratic tyranny because such strangleholds on commerce increasingly will pose both internal and external threats to the nation. We must address educational problems at their level of complexity and accept the resulting challenge to so organize ourselves that no one group can dominate the enterprise.

That professional people engage in professional activity is nine-tenths myth. Most professional people travel previously-plowed ground most of the time and so operate as technicians much more often than as professionals. Technical work can be specified although, varying in complexity, more easily in some cases than others. The notion that teachers--like it or not--will be the final arbiters of practice is understandable. They are only mimicking politicians, doctors, lawyers, professors, educational R&D personnel, and others now shaded by the umbrella of professional mystique. The technical efforts of all professionals--including teachers--should increasingly come under provisions of a social doctrine of accountability. Technical educational practice should be evaluated. Much of the effort of "big educational R&D" might usefully be devoted to providing a framework that insures that service personnel are evaluated equitably and fairly. This requirement alone destroys the classical oversimplification of evaluation as a two-stage formative-summative categorization.

Scriven's concerns with higher-order effects of education are legitimate but appear too narrowly referenced and premature excepting in the policy science sense of evaluation. We will not get any other kind of evaluation of higher-order effects until a system of social indicators such as is sketched by Bauer and his associates is developed, evaluated, and appropriately institutionalized.

Like it or not, we currently are stuck with comparative-relative evaluation for every evaluation requirement save one--first-order effects evaluation of the product and, through it, of the development staff. Entertainably, comparative-relative evaluation sometimes will be appropriate when first-order product effects require evaluation. However, one suffers this state of affairs, rather than champions it, for it typically reflects a cop-out concession to a defective educational status quo.

REFERENCES

Atkin, J. M. & Grotelueschen, A. On changing educational practice. Paper prepared for the Syracuse University Research Corporation Policy Institute. December 10, 1971.

Bauer, R.A. Detection and anticipation of impact: The nature of the task. In R. A. Bauer (Ed.), Social indicators. Cambridge, Mass.: MIT Press, 1966.

Bloom, B.S., Hastings, J.T., & Madaus, G.F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Braybrooke, D. & Lindblom, C.E. A strategy of decision. New York: Free Press, 1963.

Bronfenbrenner, U. Two worlds of childhood: U.S. and U.S.S.R. New York: Russell Sage Foundation, 1970.

Follettie, J. F. Alternative designs for educational systems. Technical Report No. 45, 1972, SWRL Educational Research and Development, Los Alamitos, California.

Jencks, C. J., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books, Inc., 1972.

Land, K. C. Social indicator models: An overview. Paper presented at AAAS Meeting, Washington, D. C., December 26, 1972.

Light, R. J. & Smith, P. V. Choosing a future: Strategies for designing and evaluating new programs. Harvard Educational Review, 1970, 40, 1-28.

Merton, R.K. Insiders and outsiders. University Lecture delivered at Columbia University, November 27, 1972. Cited in: Does it take one to know one? Columbia Reports, January 1973, p.2.

Price, D. J. de S. Little science, big science. New York: Columbia University Press, 1963.

Scriven, M. The methodology of evaluation. In R. W. Tyler, R.M. Gagne, & M. Scriven, Perspectives of curriculum evaluation. Chicago: Rand McNally, 1967.

Siegel, L., & Siegel, L. C. A multivariate paradigm for educational research. Psychological Bulletin, 1967, 68, 306-326.

Stephens, J. M. The process of schooling. New York: Holt, Rinehart & Winston, 1967.